



AUTHORITYHACKER INTERNAL LINKS STUDY 2019

Introduction

As far as I know, there are no larger scale studies on the effects of internal links on SEO. Most of what's known comes from the original PageRank paper (1996), and speculations about various things said by Google employees.

We did this study to find out if some of that knowledge is still valid.

Our Dataset

We worked with 2 different datasets. First was composed of 500k search positions, with URLs and other stats that we got out of Ahrefs Keyword Explorer.

These were the results for >1000 searches a month keywords of minimum 3 words. For each of the SERPS we obtained additional metrics using Ahrefs API.

We worked with the following key data points:

Keyword	Organic Traffic	Volume	Country	CPS	URL
Position	Backlinks	External Links	Internal Links	Referring Domains	Internal Backlinks
Linking URL	Anchors	Ahrefs URL Rating	Referring Domains	Title Tag	Content Length

Ahrefs Keyword Explorer	Ahrefs Site Explorer (Selenium Crawl)
Ahrefs API	Custom Crawler

Then, we obtained all internal backlink data (links, their metrics, anchors) for 10,000 SERPs, (~1000 for each of the top ten search results for 1000 keywords with same attributes as the first dataset.)

Building our own crawler was way too complicated as there are many variables involved that would pollute the data we could get (we tried).

Therefore, we turned to Ahrefs Site Explorer and used the 'Internal Links' analysis feature. The data were acquired in an automated fashion using a python implementation of Selenium - a web browser automation tool.

Finally, we got additional metrics using Ahrefs API and custom built crawler.

Our tech stack



Google Ads



ahrefs



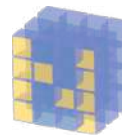
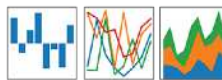
Google Cloud Platform



python™

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



NumPy



BeautifulSoup



SQLite

We relied on Google Ads API to generate random keywords, Ahrefs to acquire SERPs, Ahrefs API for additional metrics and Selenium for automating a bunch of these manual processes.

To process the data we used Python, SQLite, Pandas and Numpy.

We employed some other Python libraries, such as Readability, BeautifulSoup, re (for pattern matching), asyncio and aiohttp for custom crawler.

Our methods and analysis

This was quite a straightforward process. We ran a number of different tests based on ideas we generated by reading popular SEO articles about the topic as well as the generally known beliefs about internal linking.

We looked at the median results. However, all tests were done in multiple ways to see if there's any consistency. I.e we also looked at the mean, Spearman (where applicable), with and without potential outliers, and so on.

For some tests, we kicked out all the root domain results not to pollute the results. For example, measuring the length of the URL would be severely affected by such results.

Only valid results were published. Many assumptions did not validate, so we left them out of the article.

For more information about this study, contact AuthorityHacker or Michal Ugor at hello@mchlgr.com